Progression of Live Operation Monitoring

David Huff (New Mexico Institute of Mining and Technology)

Mentor: Daniel Illescas(HPC-ENV), Mike Mason(HPC-ENV)

Background

The Operations(Ops) team uses a Splunk Application to augment their ability to react and fix the multiple LANL clusters. Splunk is a messages analytic web application that can take in computer logs and search them for useful info. In this ever evolving field, it is important for this app to keep up with the new systems, architecture, and monitoring tools being deployed within HPC.

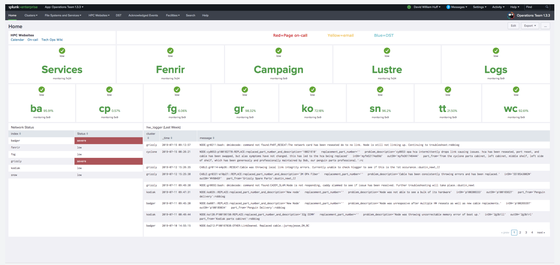


Fig. (1): This is the home dashboard the Ops team use on a daily basis

Acknowledge Script

Currently there are two Splunk search heads in production. One is for the general use of the entire HPC staff, while the second is set aside specifically for the Ops team. Although the other teams can see what errors occurred, due to the separation, they do not know what action the Ops team have taken to correct the errors. To fix this a script was implemented to sync the two databases together. The script takes the acknowledged events from a local Splunk KV Store, stored in a MongoDB, and transfers it over to an identical database on the other search head. This was done in the hopes that it would increase transiency and communication for the Ops team and cluster admins. A new Splunk cluster is also being created to improve scalability and to remove the need to sync databases.

Acknowledged Events

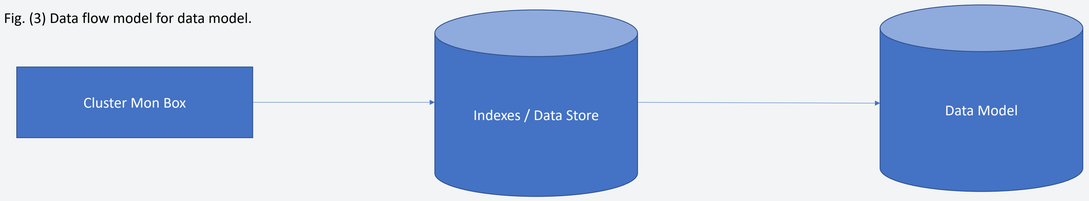
Original Event Time	Acknowledged At	Acknowledged Until	device	type	message	raw	notes	username
2019-07-11 08:29:58	2019-07-11 08:38:32	2019-07-11 13:38:32	core3 Port 26	LocalLinkIntegrityErrors	[LocalLinkIntegrityErrors == 4194397] - {edge26 Port 3}	<14>Jul 11 08:29:58 gr-master hsnmon[126901]: core3 Port 26: [LocalLinkIntegrityErrors == 4194397] - {edge26 Port 3}		
2019-07-11 08:35:22	2019-07-11 08:42:45	2019-07-12 08:42:45	gr1392	pingStatus	compute node gr1392 is down	<13>Jul 11 08:35:22 gr-mon2.lanl.gov pingStatus[111437]: compute node gr1392 is down		
2019-07-11 08:35:22	2019-07-11 08:42:45	2019-07-12 08:42:45	gr0843	pingStatus	compute node gr0843 is down	<13>Jul 11 08:35:22 gr-mon2.lanl.gov pingStatus[111436]: compute node gr0843 is down		
2019-07-11 08:29:58	2019-07-11 08:37:41	2019-07-12 08:37:41	core3 Port 26	LocalLinkIntegrityErrors	[LocalLinkIntegrityErrors == 4194397] - {edge26 Port 3}	<14>Jul 11 08:29:58 gr-master hsnmon[126901]: core3 Port 26: [LocalLinkIntegrityErrors == 4194397] - {edge26 Port 3}		
2019-07-11 09:02:24	2019-07-11 09:21:01	2019-07-12 09:21:01	gr0646 edge21-18 1M CU Missing Link	Check Cable	Check Cable: gr0646 edge21-18 1M CU Missing Link	<14>Jul 11 09:02:24 gr-master hsnmon[133977]: Check Cable: gr0646 edge21-18 1M CU Missing Link		

Fig. (2): example of what the Ops team deal with on a daily basis

Data Model Change

One of the benefits of Splunk is the ability to create schemas on the fly. This is generally beneficial when exploring your data but we found doing this repeatedly over large data sets caused performance issues. To prevent an overload, a data model was implemented to store and structure error messages explained in figure 3 . The new data model is modular, which allows for easy changes to the model when new systems are added or removed. The model also speeds up the search process allowing the Ops team better utilize their time.

Fig. (3) Data flow model for data model.



NVIDIA GPU Monitoring

With new technology comes new problems. Kodiak is a GPU cluster, however, it was found that some of the GPUs on the system were staying in a constant throttled state causing jobs to run slow. By running nvidia-smi to sample stats on Kodiak's compute nodes, we can log this data into the monitoring infrastructure. Allowing us to alert Ops when this occurs so they can take appropriate actions. The data also gives another dimension to study for future analysis.

Future Work: Statistical Analysis

Syslog

Currently we are testing out new Splunk libraries. Although, it is currently unclear if the new functionality will be implemented into a dashboard. We are using basic statistical analysis to make simple job and system profiles. The images below show both types of profiles. Figure 4 tabulates the average number of syslog messages the cluster produces every hour and the second standard deviation. This information can be used to perform regression analysis with a confidence interval of 95%. The next table compares a single job on one node against the cluster statistics to find outliers. The time charts in figure 5 are Splunk's equivalent to the regression analysis. The top time chart is the number of syslog messages produced every hour and the blue shaded area is the confidence interval.

Cooling Distribution Unit(CDU)

The bottom time chart in figure 5 is the same process but uses data collected from the CDU. With new Splunk Analytic tools we hope to improve the current CDU alerts. The current alerts often report false alarms caused by intermittent outliers from the sensors, and have difficulties detecting sensor problems vs actual CDU problems. With smarter CDU monitoring it will be easier for the Ops team to identify problems before any damage happens on the system.

lowThreshold	length	mean	highThreshold
-160.00775214955604	736	369	898.007752149556

testing for syslogs

_time	syslog_count	severity
2019-07-16 15:00	236	0
2019-07-16 16:00	204	0
2019-07-16 17:00	204	0
2019-07-16 18:00	215	0
2019-07-16 19:00	204	0
2019-07-16 20:00	204	0
2019-07-16 21:00	204	0
2019-07-16 22:00	204	0
2019-07-16 23:00	214	0
2019-07-17 00:00	213	0

« prev 1 2 next »

Fig. (4): This image depicts the the confidence interval and how we are using it to to compare syslog messages.

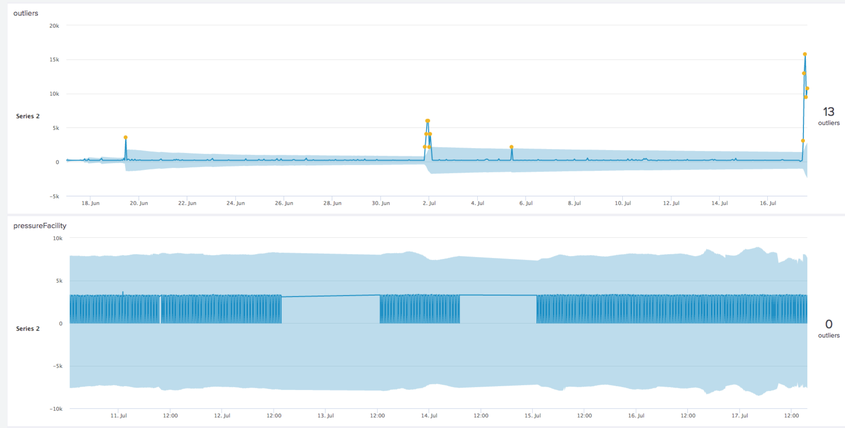


Fig. (5): The time charts are Splunk's version of regression analysis applied to syslog messages(top) and Cooling Distribution Unit(bottom)

LA-UR-19-27185